

# MODUL 8

---

## KLASIFIKASI : NAÏVE BAYES

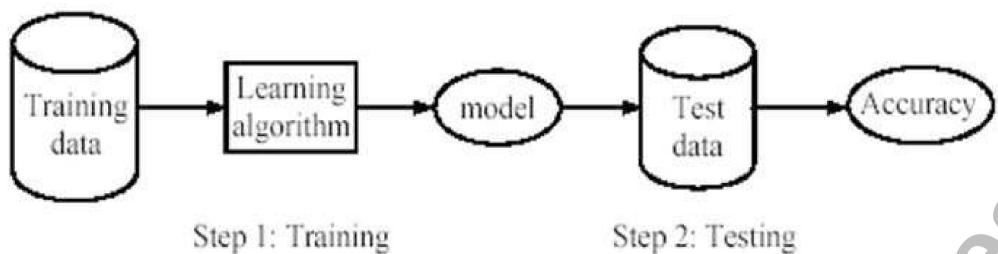
### 8.1. Tujuan

1. Mahasiswa mampu menggunakan dan membuat model klasifikasi dengan teorema Naïve Bayes.
2. Mahasiswa mampu menerapkan algoritma Naïve Bayes terhadap studi kasus tertentu.

### 8.2. Landasan Teori

Teori keputusan Bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola (*pattern recognition*). Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan resiko yang ditimbulkan dalam keputusan-keputusan tersebut.

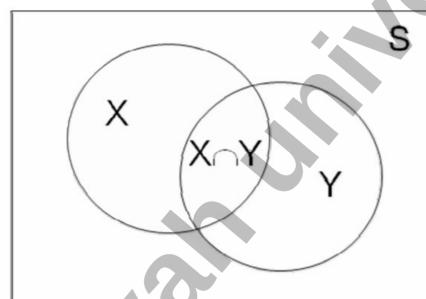
Ada dua proses penting yang dilakukan saat melakukan klasifikasi. Proses yang pertama adalah *learning (training)* yaitu proses pembelajaran menggunakan *training set*. Untuk kasus Naïve Bayesian Classifier, perhitungan probabilitas dari data berdasarkan data pembelajaran dilakukan. Proses yang kedua adalah proses *testing* yaitu menguji model menggunakan *data testing*. Gambar berikut memperlihatkan alur dari kedua proses tersebut.



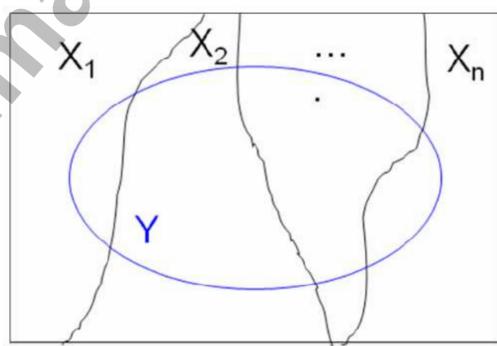
**Gambar 8.1** Tahapan Proses Klasifikasi

Metode Bayes menggunakan probabilitas bersyarat sebagai dasarnya. Dalam ilmu probabilitas bersyarat dinyatakan sebagai:

$$P(X | Y) = \frac{P(X \cap Y)}{P(Y)}$$



Metode Bayes dan HMAP (*Hypothesis Maximum Appropri Probability*)



$$P(X_k | Y) = \frac{P(Y | X_k)}{\sum_i P(Y | X_i)}$$

Dimana keadaan Posteriror (Probabilitas  $X_k$  di dalam  $Y$ ) dapat dihitung dari keadaan prior (Probabilitas  $Y$  di dalam  $X_k$  dibagi dengan jumlah probabilitas  $Y$  dalam semua  $X_i$ ).

Terminologi dari HMAP menyatakan hipotesa yang diambil berdasarkan nilai probabilitas berdasarkan kondisi prior yang diketahui.

HMAP adalah model penyederhanaan dari metode bayes yang disebut dengan Naive Bayes. HMAP dapat digunakan sebagai metode untuk mendapatkan hipotesis dari suatu keputusan. HMAP dapat diartikan untuk mencari probabilitas terbesar dari semua instance pada atribut target atau semua kemungkinan keputusan.

Salah satu manfaat algoritma naïve bayes adalah untuk melakukan prediksi terhadap data-data tertentu. Prediksi (testing) terhadap data yang akan datang bisa dilakukan berdasarkan hasil pembelajaran terhadap data training. Data training diambil dari data yang terdahulu, sedangkan data uji (testing) bisa diambil dari data-data yang sedang atau akan terjadi.

### 8.3. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Ms. Excel, Weka, RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

### 8.4. Langkah-langkah Praktikum

#### 8.4.1. Implementasi Naïve Bayes dengan Weka

Langkah-langkah menggunakan algoritma naïve bayes dengan Weka sebagai berikut:

1. Persiapkan file **Cuaca.arff** dari hasil percobaan kegiatan 7.4.1 pada Modul 7. File ini akan kita gunakan sebagai data training.
2. Buatlah sebuah data testing dengan format **ARFF** dari tabel

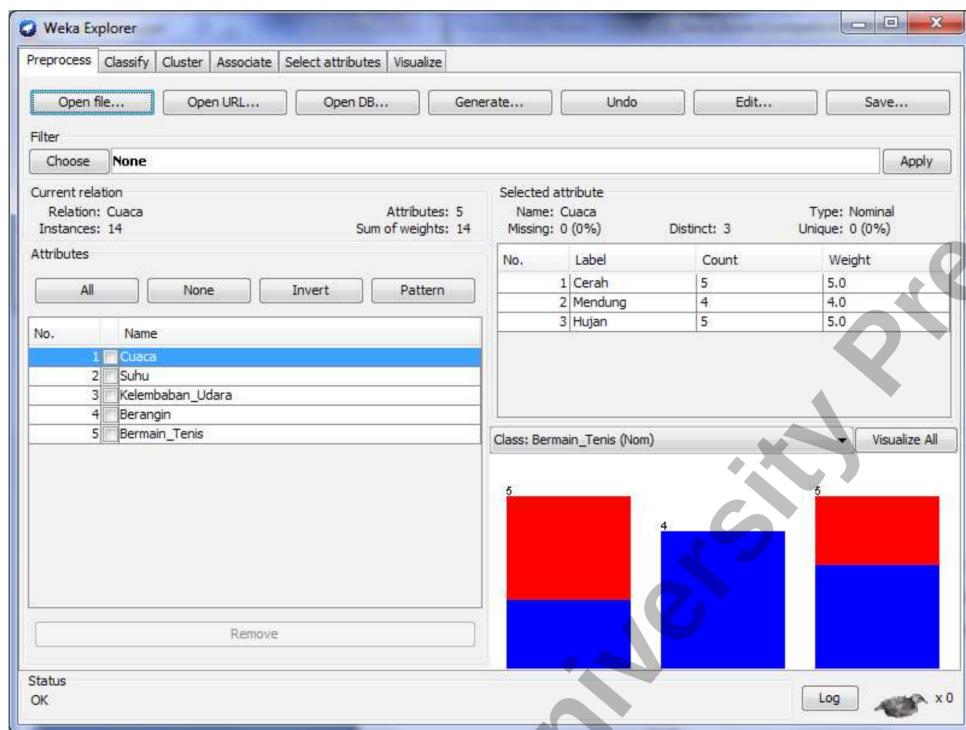
8.1 sebagai data uji yang akan diprediksi dengan memiliki variabel-variabel independen dan variabel dependen yang sama. Dengan ketentuan variabel dependen diisi dengan tanda tanya (?). Asumsi bahwa kita belum mengetahui nilai / kelas dari variabel tersebut. Nilai / kelas inilah yang akan kita prediksi dengan menggunakan algoritma naïve bayes.

3. Simpan dengan nama **CuacaTesting.arff**.

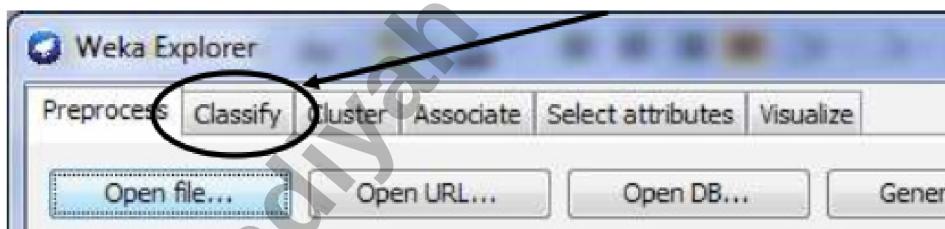
Tabel 8.1 Data Testing Cuaca

Cuaca	Suhu	Kelembaban_udara	Berangin	Bermain_Tenis
Cerah	75	65	TIDAK	?
Cerah	80	68	YA	?
Cerah	83	87	YA	?
Mendung	70	96	TIDAK	?
Mendung	68	81	TIDAK	?
Hujan	65	75	YA	?
Hujan	64	85	YA	?

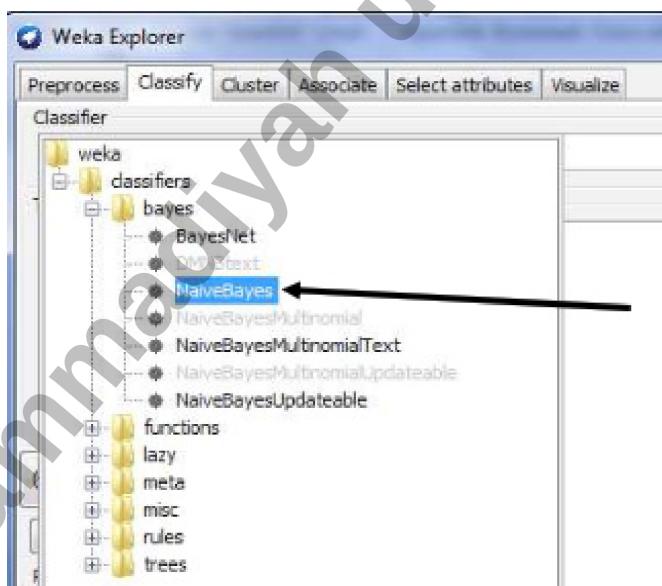
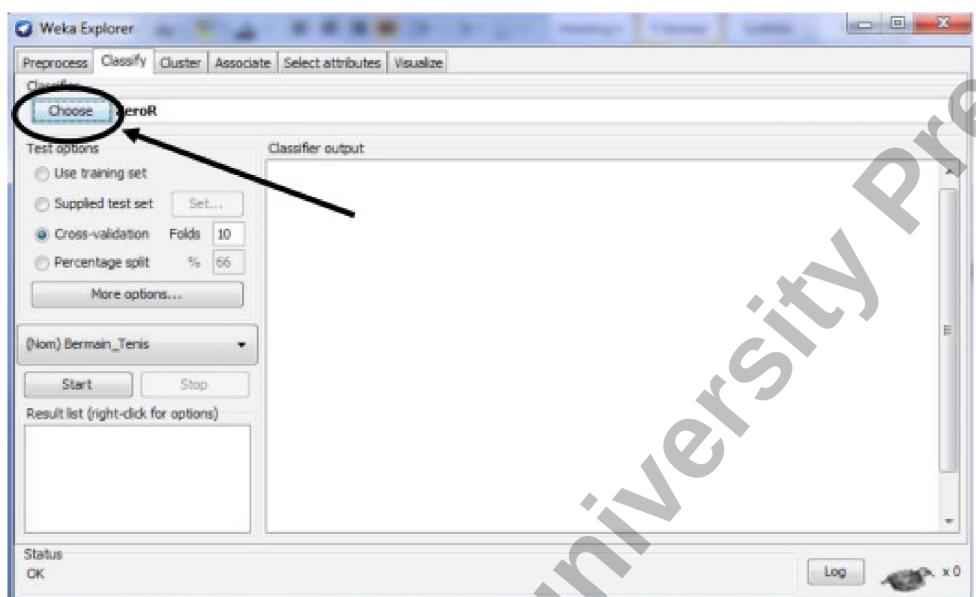
4. Jika telah selesai membuat Buka aplikasi Weka, masuk dalam menu Weka Explorer.
5. Buka kembali file **Cuaca.arff** dari hasil kegiatan 7.4.1 pada Modul 7 dengan menggunakan Weka Explorer. File ini akan kita gunakan sebagai data pelatihan untuk memprediksi data testing pada file **CuacaTesting.arff**.



6. Masih pada jendela Weka Explorer, pilih tab **Classify**.



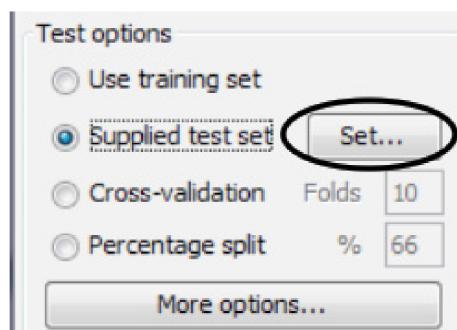
7. Sehingga akan muncul jendela Weka Explorer pada tab Classify. Pada kotak **Classifier** klik tombol **Choose** untuk memilih metode / algoritma **Naïve Bayes**.



8. Selanjutnya adalah menentukan data testing sebagai data yang akan diprediksi variabel dependennya. File **CuacaTesting.arff** ditentukan sebagai data testing pada kegiatan ini.

9. Pada menu Test Options terdapat 4 pilihan pengujian, yaitu:
- Use training set, jika data pelatihan dan data uji menggunakan file ARFF yang sama. Pada pilihan ini, data yang akan diprediksi menggunakan data training.
  - Supplied test set, jika data uji telah disediakan dalam file ARFF yang lain terpisah dengan data training.
  - Cross-validation, nilai default Folds = 10. Hal ini berarti sistem akan mengacak data training set dan mengambil sebagian dari datanya untuk dijadikan testing set. Proses ini dilakukan sebanyak 10 kali dan hasil akhir merupakan akurasi rata-rata dari sepuluh percobaan tersebut.
  - Percentage split, dengan nilai default 66%. Hal ini berarti sistem akan mengambil sebanyak 66% dari seluruh data yang ada sebagai data pelatihan dan sisanya digunakan sebagai data uji.
10. Pada percobaan kali ini, kita akan menggunakan pilihan **Supplied test set**.

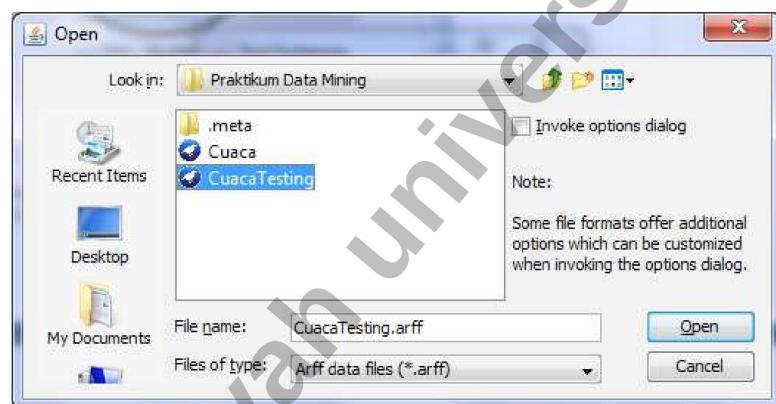
Klik tombol Set untuk menentukan file ARFF sebagai data uji.



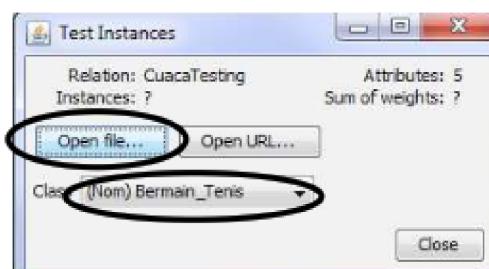
11. Sehingga akan muncul jendela Test Instance. Klik **Open file...**



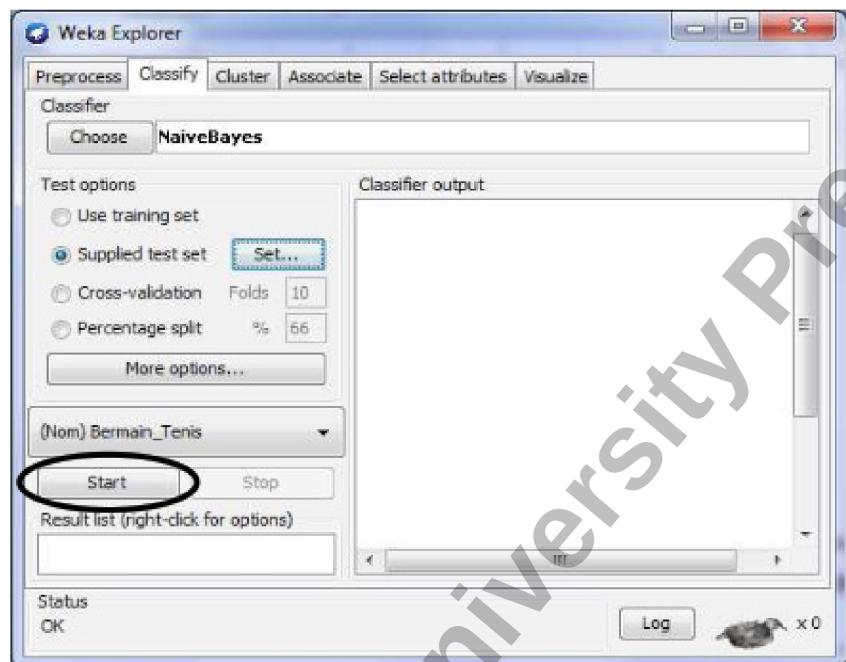
12. Pilih file **CuacaTesting.arff** sebagai data uji. Klik **Open**.



13. File **CuacaTesting.arff** akan diset sebagai data uji pada jendela Test Instances dengan variabel predictor (Class) adalah Bermain\_Tenis. Klik **Close**.

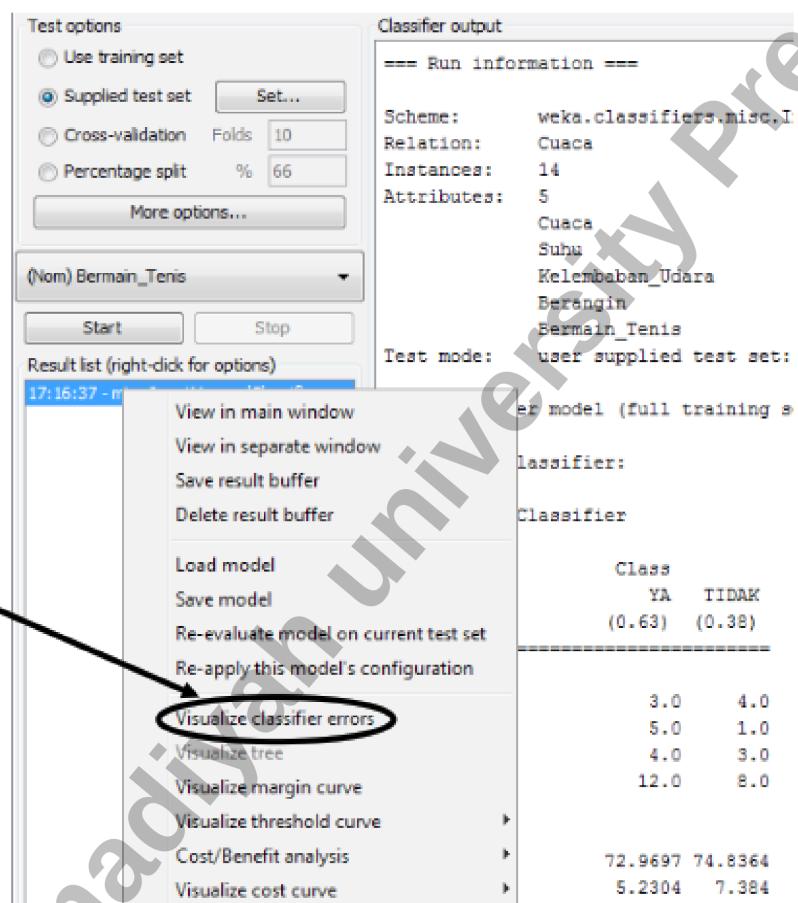


14. Klik **Start** untuk memulai proses naïve bayes.

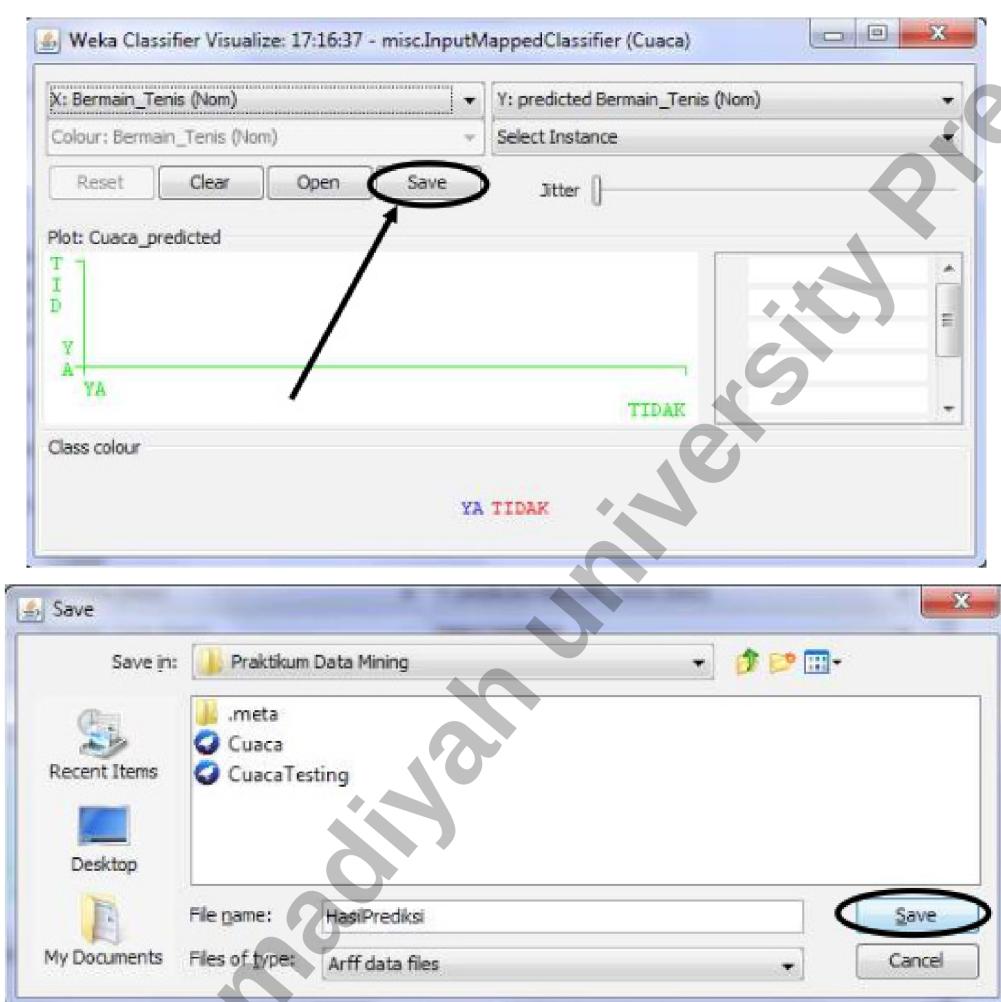


15. Jika muncul jendela pesan **Classifier Panel**, kita abaikan dengan mengklik Yes. Sehingga algoritma naïve bayes akan diproses.
16. Karena pada percobaan ini kita memproses data uji yang belum diketahui nilai kelas dari variabel dependen yang diajukan, maka kita abaikan nilai-nilai yang ditampilkan dalam jendela **Classifier Output**.

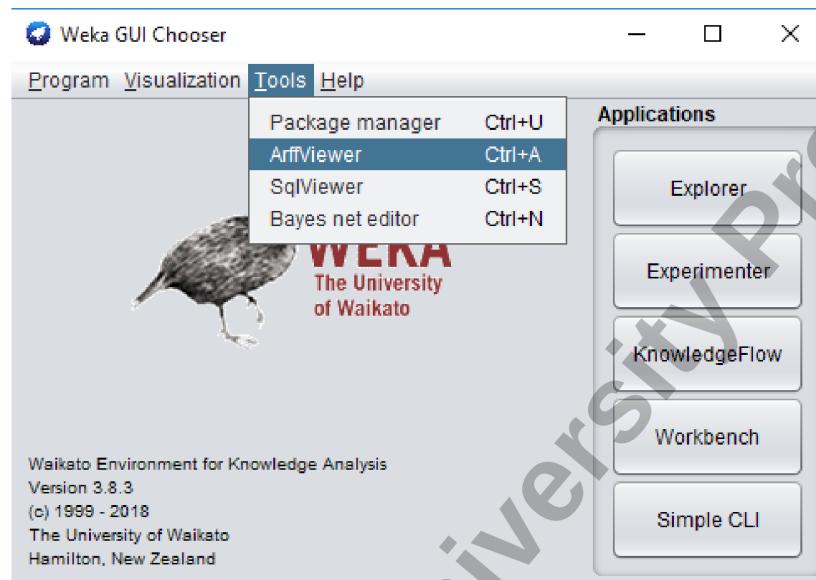
17. Untuk melihat hasil prediksi terhadap data uji, yang perlu kita lakukan berikutnya adalah dengan melihat nilai **Classifier Errors**. Klik kanan pada hasil proses dalam kotak **result list**. Pilih menu **Visualize classifier errors**.



18. Pada jendela Weka Classifier Visualize, abaikan hasil apapun yang ditampilkan. Klik **Save**. Simpan dengan nama file **HasilPrediksi.arff**.



19. Tutup semua jendela termasuk Weka Explorer dan kembali ke **Weka GUI Chooser**. Pilih menu **Tools – ArffViewer**.



20. Jendela ARFF-Viewer akan ditampilkan. Buka menu **File – Open**. Tunjukkan pada file **HasilPrediksi.arff** yang telah anda simpan pada langkah ke-18. Lihatlah, hasil prediksi telah diketahui pada kolom **predicted Bermain\_Tenis Nominal**.

No.	1: Cuaca Nominal	2: Suhu Numeric	3: Kelembaban_Udara Numeric	4: Berangin Nominal	5: prediction margin Numeric	6: predicted Bermain_Tenis Nominal
1	Cerah	75.0	65.0	TIDAK	0.88138	YA
2	Cerah	80.0	68.0	YA	0.54393	YA
3	Cerah	83.0	87.0	YA	0.16156	TIDAK
4	Mendung	70.0	96.0	TIDAK	0.81426	YA
5	Mendung	68.0	81.0	TIDAK	0.91699	YA
6	Hujan	65.0	75.0	YA	0.62686	YA
7	Hujan	64.0	85.0	YA	0.41992	TIDAK

### 8.4.2. Implementasi Naïve Bayes dengan RapidMiner

Penggunaan algoritma naïve bayes dengan RapidMiner untuk melakukan prediksi pada dasarnya sama dengan menggunakan Weka. Kita perlu mempersiapkan data training dan data testing. Bedanya terletak pada format file yang digunakan. Jika dengan Weka file yang digunakan memiliki format ARFF, sedangkan jika menggunakan RapidMiner bisa dilakukan terhadap file excel.

Berikut langkah-langkahnya:

1. Persiapkan file **Tabel\_Cuaca.xls** yang terdiri dari 2 sheet.
2. Sheet1 digunakan sebagai data training, dan sheet2 digunakan sebagai data uji.
3. Masing-masing tabel memiliki atribut yang sama, yaitu:
  - a) Cuaca (X1)
  - b) Suhu (X2)
  - c) Kelembaban\_udara (X3)
  - d) Berangin (X4)
  - e) Bermain\_Tenis (Y), sebagai variabel predictor.

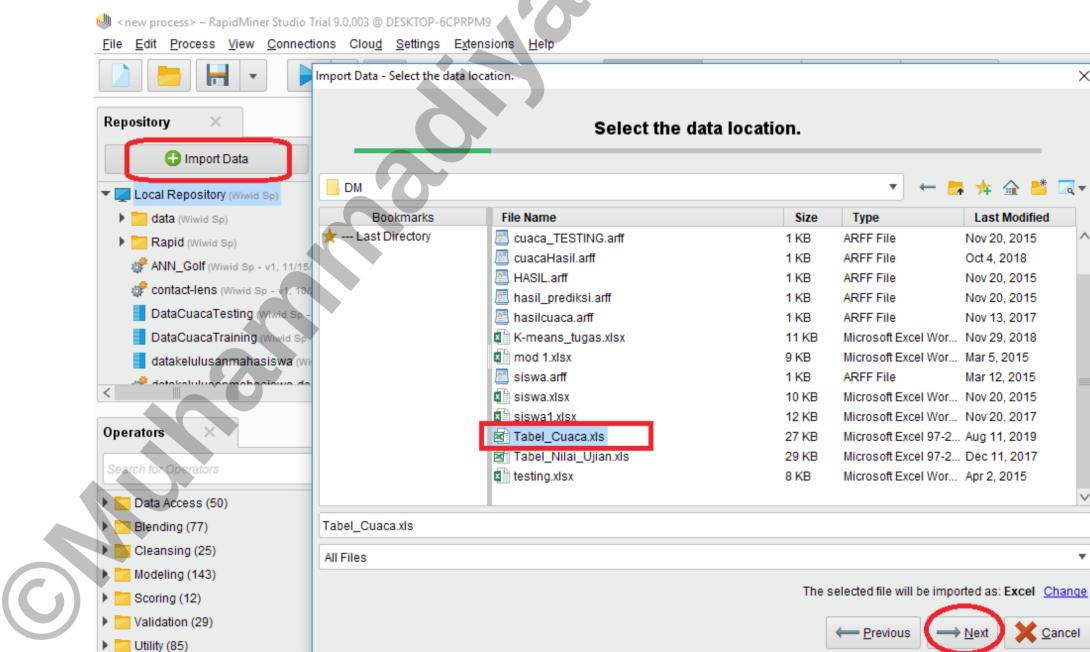
**Tabel data training** pada Sheet1

	A	B	C	D	E
1	Cuaca	Suhu	Kelembaban_udara	Berangin	Bermain_Tenis
2	Cerah	85	85	TIDAK	TIDAK
3	Cerah	80	90	YA	TIDAK
4	Mendung	83	86	TIDAK	YA
5	Hujan	70	96	TIDAK	YA
6	Hujan	68	80	TIDAK	YA
7	Hujan	65	70	YA	TIDAK
8	Mendung	64	65	YA	YA
9	Cerah	72	95	TIDAK	TIDAK
10	Cerah	69	70	TIDAK	YA
11	Hujan	75	80	TIDAK	YA
12	Cerah	75	70	YA	YA
13	Mendung	72	90	YA	YA
14	Mendung	81	75	TIDAK	YA
15	Hujan	71	91	YA	TIDAK

**Tabel data uji** pada Sheet2 tanpa ada variabel **Bermain\_Tenis**.

	A	B	C	D
1	Cuaca	Suhu	Kelembaban_udara	Berangin
2	Cerah	75	65	TIDAK
3	Cerah	80	68	YA
4	Cerah	83	87	YA
5	Mendung	70	96	TIDAK
6	Mendung	68	81	TIDAK
7	Hujan	65	75	YA
8	Hujan	64	85	YA

4. Buka aplikasi **RapidMiner**. Menjalankan RapidMiner untuk pertama kali, terlebih dahulu membuat repositori baru. Repositori ini berfungsi sebagai lokasi penyimpanan terpusat untuk data dan proses analisa.
5. Klik **Import Data**. Arahkan direktori tempat penyimpanan file pada langkah **Select the data location**, kemudian pilih file yang akan digunakan dan klik **Next**.



6. Pastikan sel Excel sesuai di langkah **Select the cells to import.**

Import Data - Select the cells to import.

**Select the cells to import.**

Sheet: Training Cell range: A:E Select All Define header row: 1

	A	B	C	D	E
1	Cuaca	Suhu	Kelembaban_udara	berangin	Bermain_Tenis
2	Cerah	85	85	TIDAK	TIDAK
3	Cerah	80	90	YA	TIDAK
4	Mendung	83	86	TIDAK	YA
5	Hujan	70	96	TIDAK	YA
6	Hujan	68	80	TIDAK	YA
7	Hujan	65	70	YA	TIDAK
8	Mendung	64	65	YA	YA
9	Cerah	72	95	TIDAK	TIDAK
10	Cerah	69	70	TIDAK	YA
11	Hujan	75	80	TIDAK	YA
12	Cerah	75	70	YA	YA
13	Mendung	72	90	YA	YA
14	Mendung	81	75	TIDAK	YA
15					

← Previous → Next Cancel

7. Pada langkah **Format your columns** ubah kolom **Bermain\_Tenis** dengan tipe data **binomial** karena hanya ada dua keputusan (YA dan TIDAK).

Import Data - Format your columns.

**Format your columns.**

Date format: MMM d, yyyy h:mm:ss a z Replace errors with missing values

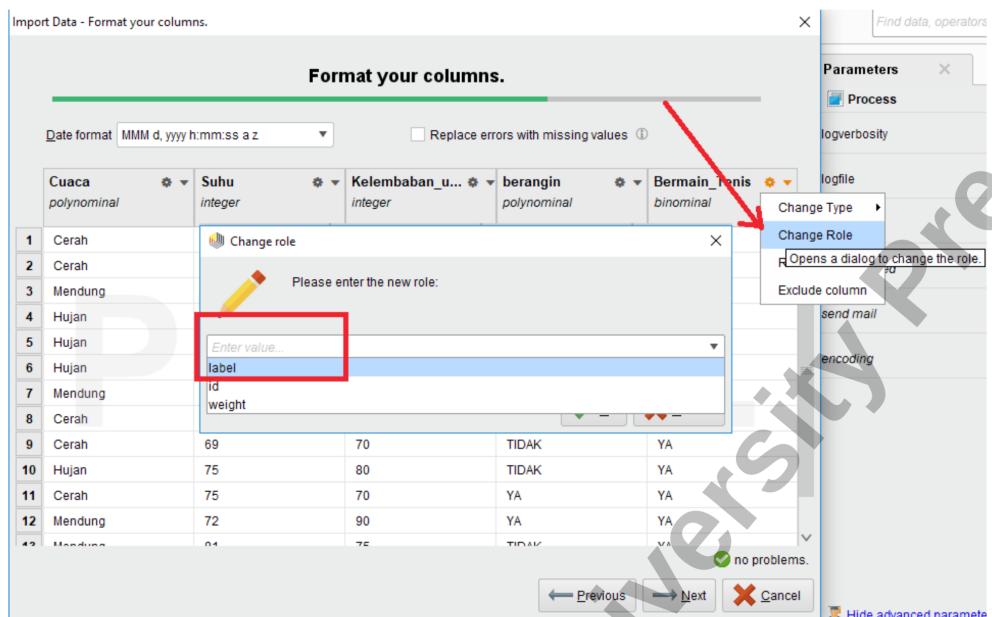
Cuaca	Suhu	Kelembaban_u...	berangin	Bermain_Tenis
polynomial	integer	integer	polynomial	polynomial
1 Cerah	85	85	TIDAK	TIDAK
2 Cerah	80	90	YA	TIDAK
3 Mendung	83	86	TIDAK	YA
4 Hujan	70	96	TIDAK	YA
5 Hujan	68	80	TIDAK	YA
6 Hujan	65	70	YA	TIDAK
7 Mendung	64	65	YA	YA
8 Cerah	72	95	TIDAK	TIDAK
9 Cerah	69	70	TIDAK	YA
10 Hujan	75	80	TIDAK	YA
11 Cerah	75	70	YA	YA
12 Mendung	72	90	YA	YA
13 Mendung	81	75	TIDAK	YA

Change Type: polynominal binomial real integer date\_time

no problems.

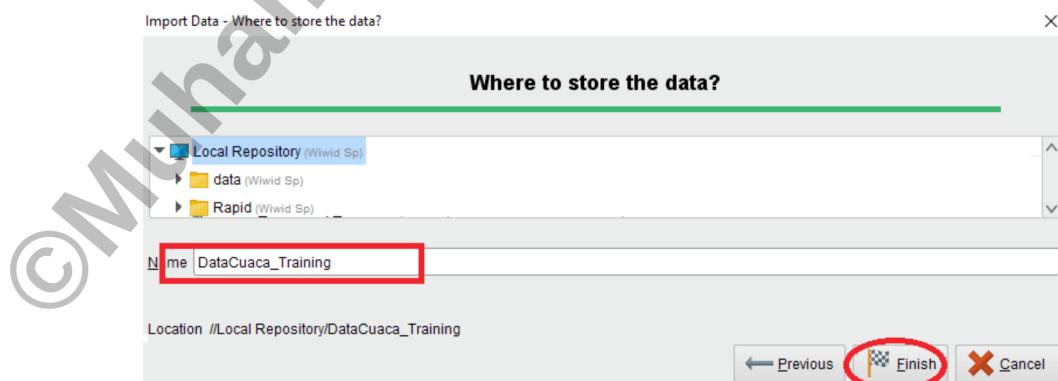
← Previous → Next Cancel Hide advanced parameters

8. Ubah pula sebagai **label** pada **Change Role**.



	Cuaca	Suhu	Kelembaban_u...	berangin	Bermain_Tenis
	polynomial	integer	integer	polynomial	binomial
1	Cerah				
2	Cerah				
3	Mendung				
4	Hujan				
5	Hujan				
6	Hujan				
7	Mendung				
8	Cerah	69	70	TIDAK	YA
9	Cerah	75	80	TIDAK	YA
10	Hujan	75	70	YA	YA
11	Cerah	72	90	YA	YA
12	Mendung	84	75	TIDAK	YA
13	Mendung				

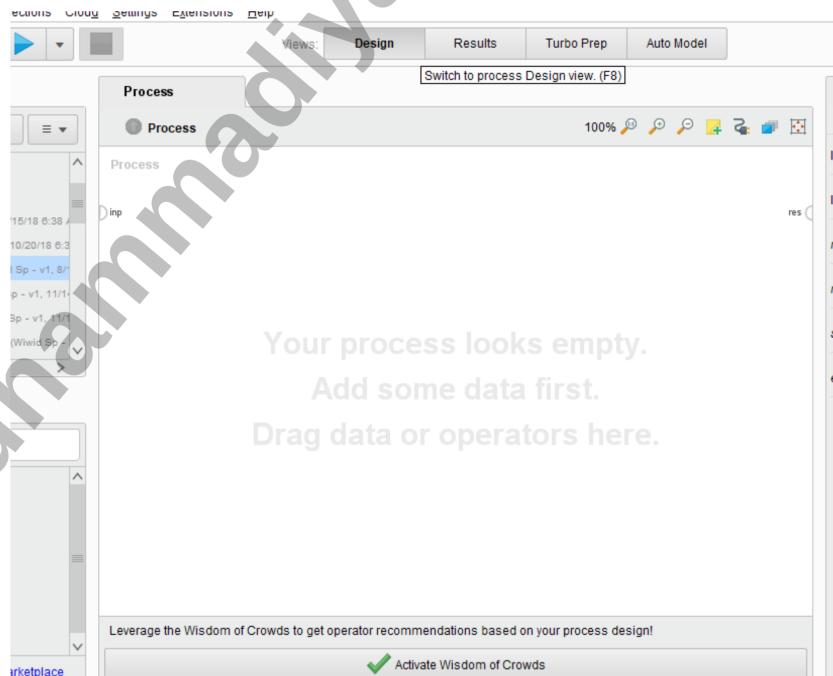
9. Simpan dengan nama **DataCuaca\_Training** dilanjutkan klik tombol **Finish**.



10. Hasil import file **Tabel\_Cuaca.xls** pada Sheet1 akan ditampilkan.

Row No.	Bermain_Tenis	Cuaca	Suhu	Kelembaban_udara	berangin
1	TIDAK	Cerah	85	85	TIDAK
2	TIDAK	Cerah	80	90	YA
3	YA	Mendung	83	86	TIDAK
4	YA	Hujan	70	96	TIDAK
5	YA	Hujan	68	80	TIDAK
6	TIDAK	Hujan	65	70	YA
7	YA	Mendung	64	65	YA
8	TIDAK	Cerah	72	95	TIDAK
9	YA	Cerah	69	70	TIDAK
10	YA	Hujan	75	80	TIDAK
11	YA	Cerah	75	70	YA
12	YA	Mendung	72	90	YA
13	YA	Mendung	81	75	TIDAK
14	TIDAK	Hujan	71	91	YA

11. Kembali ke jendela Design Perspective dengan shortcut tombol F8.



12. Lakukan hal yang sama untuk data testing yang diambil dari **Tabel\_Cuaca.xls** pada Sheet2 (Testing) dengan mengulang langkah 5. Pastikan semua variabel data testing terpilih (ada 4 variabel), bedanya pada langkah ini tidak ada variabel yang diubah bertipe **label** seperti pada langkah 8.

Import Data - Select the cells to import.

Select the cells to import.

Sheet: Testing ▾ Cell range: A1:D8 Select All  Define header row: 1

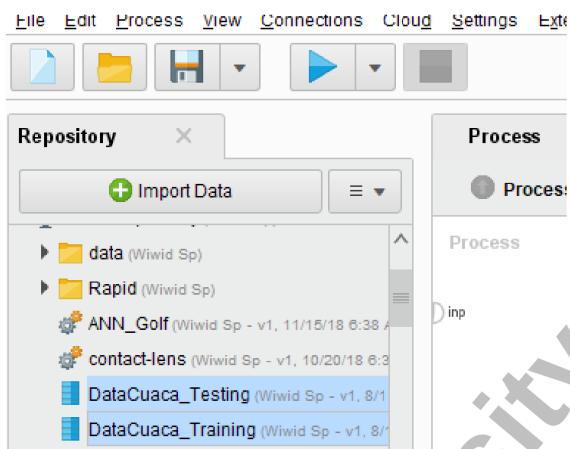
	A	B	C	D	E
1	<b>Cuaca</b>	<b>Suhu</b>	<b>Kelembaban_udara</b>	<b>berangin</b>	<b>Bermain_Tenis</b>
2	Cerah	75	65	TIDAK	
3	Cerah	80	68	YA	
4	Cerah	83	87	YA	
5	Mendung	70	96	TIDAK	
6	Mendung	68	81	TIDAK	
7	Hujan	65	75	YA	
8	Hujan	64	85	YA	

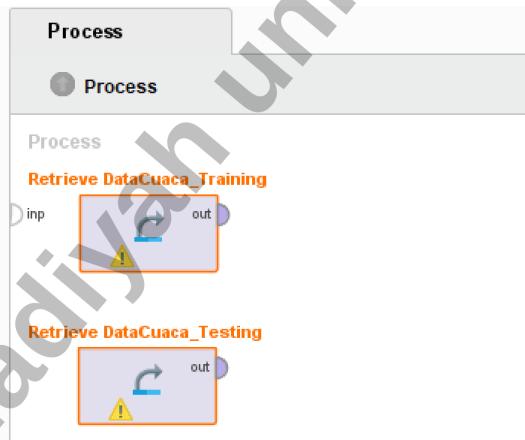
ExampleSet (7 examples, 0 special attributes, 4 regular attributes)

Row No.	Cuaca	Suhu	Kelembaban_udara	berangin
1	Cerah	75	65	TIDAK
2	Cerah	80	68	YA
3	Cerah	83	87	YA
4	Mendung	70	96	TIDAK
5	Mendung	68	81	TIDAK
6	Hujan	65	75	YA
7	Hujan	64	85	YA

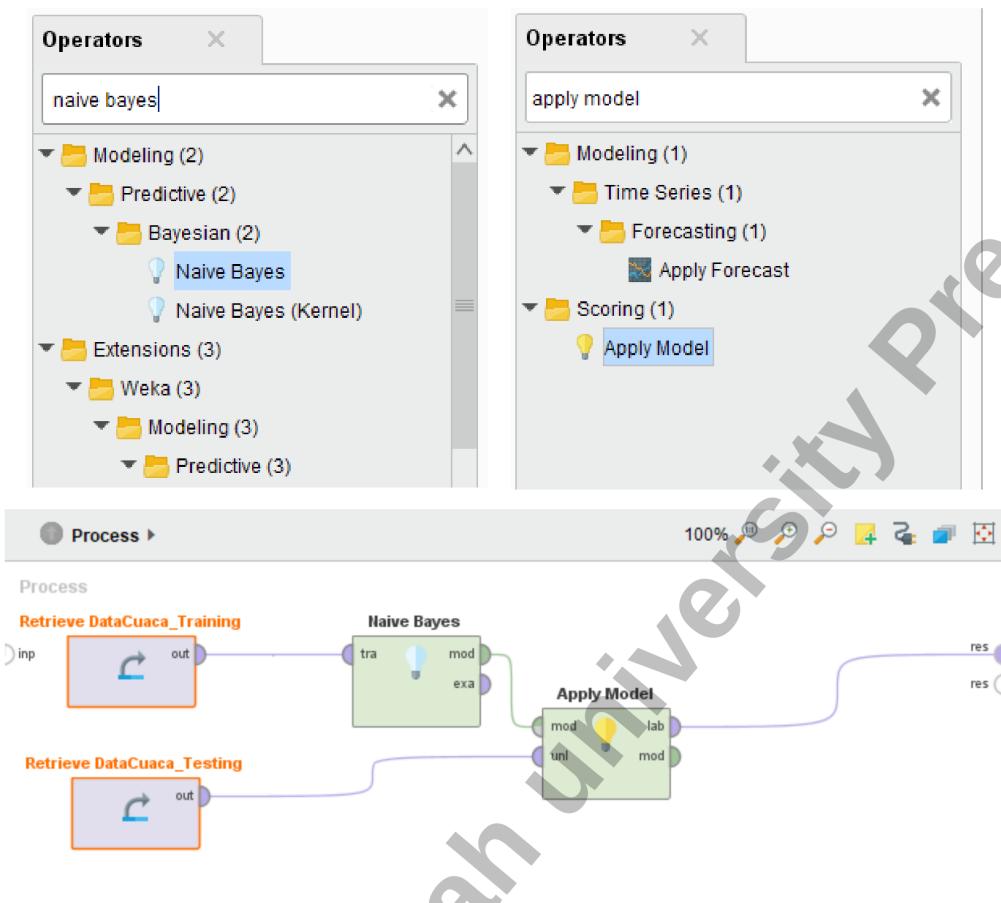
13. Simpan dengan nama **DataCuaca\_Testing**.



14. Langkah selanjutnya adalah membuat desain Naïve Bayes. Drag **DataCuaca\_Training** dan **DataCuaca\_Testing** ke dalam jendela Process View.



15. Masukkan juga operator **Naïve Bayes** dan **Apply Model** ke dalam Process View. Hubungkan konektor masing-masing data terhadap operator seperti gambar.



16. Jalankan proses naïve bayes dengan menekan tombol **Run** (atau menekan tombol F11).
17. Perhatikan hasil proses klasifikasi naïve bayes. Pada tab **Data**, dapat dilihat hasil prediksi terhadap data testing serta tingkat confidence nilai kelas pada masing-masing data.

Row No.	prediction(B...)	confidence(TIDAK)	confidence(YA)	Cuaca	Suhu	Kelembaban...	berangin
1	YA	0.154	0.846	Cerah	75	65	TIDAK
2	YA	0.498	0.502	Cerah	80	68	YA
3	TIDAK	0.856	0.144	Cerah	83	87	YA
4	YA	0.019	0.981	Mendung	70	96	TIDAK
5	YA	0.007	0.993	Mendung	68	81	TIDAK
6	YA	0.371	0.629	Hujan	65	75	YA
7	TIDAK	0.568	0.432	Hujan	64	85	YA

Pada tab **Statistics**, dapat dilihat bahwa distribusi nilai kelas pada variabel Y (Bermain\_Tenis) rerata nilai confidence sebesar 0,353 untuk nilai TIDAK, dan 0,647 untuk nilai YA.

	Name	Type	Missing	Statistics	Filter (7 / 7 attributes):
	<b>Prediction</b> <b>prediction(Bermain_Tenis)</b>	Binominal	0	Least TIDAK (2) Most YA (5) Values YA (5), TIDAK (2)	
	<b>Confidence_TIDAK</b> <b>confidence(TIDAK)</b>	Real	0	Min 0.007 Max 0.856 Average 0.353	
	<b>Confidence_YA</b> <b>confidence(YA)</b>	Real	0	Min 0.144 Max 0.993 Average 0.647	
	<b>Cuaca</b>	Polynominal	0	Least Mendung (2) Most Cerah (3) Values Cerah (3), Hujan (2), ...	
	<b>Suhu</b>	Integer	0	Min 64 Max 83 Average 72.143	
	<b>Kelembaban_udara</b>	Integer	0	Min 65 Max 96 Average 79.571	
	<b>berangin</b>	Polynominal	0	Least TIDAK (3) Most YA (4) Values YA (4), TIDAK (3)	

18. Bandingkan dengan hasil prediksi menggunakan WEKA. Dapat dilihat bahwa prediksi masing-masing aplikasi menunjukkan hasil yang sama.

## 8.5. Tugas

1. Berdasarkan tabel berikut, buatlah file dalam format Excel (.xls) dan format ARFF (.arff) ! Data ini akan digunakan sebagai **data testing**.

Jurusan_SMA	Gender	Asal_Sekolah	Rerata_SKS	Asisten
LAIN	WANITA	SURAKARTA	18	TIDAK
IPA	PRIA	SURAKARTA	19	YA
LAIN	PRIA	SURAKARTA	19	TIDAK
IPS	PRIA	LUAR	17	TIDAK
LAIN	WANITA	SURAKARTA	17	TIDAK
IPA	WANITA	LUAR	18	YA
IPA	PRIA	SURAKARTA	18	TIDAK
IPA	PRIA	SURAKARTA	19	TIDAK
IPS	PRIA	LUAR	18	TIDAK
LAIN	WANITA	SURAKARTA	18	TIDAK

2. Gunakan file ARFF yang dikerjakan pada Tugas nomor 1 dalam Modul 7 sebagai data training. Lakukan prediksi terhadap data testing (ARFF) di atas menggunakan WEKA !
3. Gunakan file Excel yang dikerjakan pada Tugas nomor 1 dalam Modul 6 sebagai data training. Lakukan prediksi terhadap data testing (Excel) di atas menggunakan RapidMiner !
4. Dari hasil percobaan Tugas nomor 3 di atas, berapakah nilai rerata confidence untuk atribut Lama\_studi dengan nilai TEPAT? Berapakah nilai rerata confidence untuk atribut Lama\_studi dengan nilai TERLAMBAT?
5. Dari hasil percobaan Tugas nomor 3 di atas, berapa orang yang akan lulus TEPAT, dan berapa orang yang akan lulus TERLAMBAT?  
Tambahkan 2 kondisi berikut pada data testing.
6. Prediksikan ketepatan lama studi si Dewi, jika Dewi adalah seorang WANITA yang berasal dari jurusan IPA pada saat SMA, asal sekolah dari LUAR SURAKARTA, mengambil SKS dengan rata-rata sebanyak 18 SKS tiap semester, dan tidak pernah menjadi Asisten selama kuliah.
7. Prediksikan ketepatan lama studi si Jono, jika Jono adalah seorang PRIA yang berasal dari jurusan selain IPA dan IPS pada saat SMA, asal sekolah dari SURAKARTA, mengambil SKS dengan rata-rata sebanyak 17 SKS tiap semester, dan pernah menjadi Asisten selama kuliah.